CLAIMS

1. A method for identifying sequences of molecules and sequence modifications from mass spectrometry data comprising:

a. producing at least one *de novo* sequence from mass spectrometry data of sequences of molecules,

b. calculating at least one mass-based alignment between each *de novo* sequence and sequences in a sequence database, wherein the molecular masses of molecules in the *de novo* sequence are compared to molecular masses of molecules in each sequence in the sequence database,

c. interpreting mass differences of modification sites between the sequence in the sequence database and the *de novo* sequence that have been identified by the mass-based alignment as modifications identified in a modification catalog,

d. calculating at least one match score for the mass-based alignment that provides an indication of matching between the sequence in the sequence database and the *de novo* sequence,

e. identifying sequences in the sequence database from mass-based alignments in response to the match scores, and

f. grouping identifications of sequences in the sequence database from at least one *de novo* sequence into an identified macromolecule list that agrees with the de novo sequencing results.

2. The method of claim 1, wherein the mass spectrometry data is generated from a tandem mass spectrometer device.

3. The method of claim 1, wherein at least one *de novo* sequence is an estimated sequence of molecules generated from the mass spectrometry data derived from a sequence of molecules.

4. The method of claim 3, wherein a *de novo* sequence is a complete or partial sequence of molecules.

5. The method of claim 3, wherein a *de novo* sequence contains incorrect or unidentifiable region of molecules where the exact sequence of molecules cannot be determined.

6. The method of claim 5, wherein a mass region is the molecular mass of the molecules in an unidentifiable region of molecules.

7. The method of claim 1, wherein at least one molecule is an amino acid and at least one sequence of molecules is a peptide.

8. The method of claim 7, wherein the peptides are derived by an enzymatic digestion of proteins.

9. The method of claim 7, wherein the sequence database is a database of amino acid sequences of proteins.

10. The method of claim 7, wherein the sequence database is a database of amino acid sequences derived from nucleotide sequences.

11. The method of claim 7, wherein the sequence database is a database of *de novo* peptide sequences.

12. The method of claim 7, wherein the sequence in the sequence database is a particular amino acid sequence in the sequence database.

13. The method of claim 6, further comprising:

a. identifying a sequence in the sequence database with a tag match, and

b. generating a mass-based alignment between a *de novo* sequence and the sequence in the sequence database.

14. The method of claim 13, wherein a mass-based alignment is a series of consecutive local mass-based alignments on either side of a tag match.

15. The method of claim 14, wherein a tag match is when a tag in the *de novo* sequence has been shown to be equivalent to a tag in a sequence in the sequence database by way of a tag search.

16. The method of claim 15, wherein a tag search is used to identify a subset of sequences in the sequence database from which to compute mass-based alignments.

17. The method of claim 16, wherein a tag is a sequence of consecutive molecules of a specified length, and the specified length is 2 to 4 molecules in length.

18. The method of claim 16, wherein single molecules of the tag and sequences in the sequence database that have the same nominal weight are represented by a single molecule.

19. The method of claim 14, wherein molecules at either side of the tag match in both the *de novo* sequence and the sequence of the sequence database are converted into mass objects.

20. The method of claim 19, wherein a mass object is at least one molecular mass and a name for that mass.

21. The method of claim 18, wherein for single molecules, mass objects are assigned the molecular mass of the single molecule.

22. The method of claim 18, wherein for unidentifiable mass regions, mass objects are assigned the molecular mass of the unidentifiable mass region.

23. The method of claim 18, wherein for reference amino acids, which represent multiple amino acids, mass objects are assigned the molecular mass of each amino acid.

24. The method of claim 19, wherein for variably modified amino acids, mass objects are assigned multiple molecular masses.

25. The method of claim 19, wherein mass regions are treated as single molecules with a single molecular mass.

26. The method of claim 19, wherein a gap is a mass object of zero molecular mass that represents no molecule.

27. The method of claim 19, wherein a local mass-based alignment is a matching of at least one consecutive mass object in the sequence in the sequence database and at least one consecutive mass object in a *de novo* sequence.

28. The method of claim 27, wherein each local mass-based alignment is generated with a breadth-first search, wherein all possible sequential combinations of mass objects of the next specified number of mass objects are compared.

29. The method of claim 28, wherein the specified number of mass objects used in the breadth first search is the search depth.

30. The method of claim 29, wherein the search depth is 3-5.

31. The method of claim 21, wherein the breadth first search is used identify the local mass-based alignment as either a mass match, a substitution, or a gap match:

32. The method of claim 31, wherein the breadth first search first tries to identify a mass match, as a local mass-based alignment where the sum of the molecular masses of the consecutive mass objects

in the sequence in the sequence database and the sum of the molecular masses of the consecutive mass objects in a *de novo* sequence are equal within a specified mass tolerance.

33. The method of claim 31, wherein if there are no mass objects left on the side of the tag match in the sequence in the sequence database, a gap match is identified as a local mass-based alignment between a gap and at least one consecutive mass object in either the sequence in the sequence database or the *de novo* sequence.

34. The method of claim 31, wherein if a mass match or a gap cannot be identified, then the breadth first search identifies a modification site as a local mass-based alignment where the sum of the molecular masses of the consecutive mass objects in the sequence in the sequence database and the sum of the molecular masses of the consecutive mass objects in a *de novo* sequence are not equal within a specified mass tolerance.

35. The method of claim 31, wherein the number of mass objects in the *de novo* sequence and the number of mass objects in the sequence database is minimized.

36. The method of claim 31, wherein the specified mass tolerance is designated by a mass tolerance of a tandem mass spectrometer device that generates the mass spectrometry data.

37. The method of claim 28, wherein a new local mass-based alignment is generated starting from the next molecule in the *de novo* sequence and the next molecule in the sequence in the sequence database after the last molecule that is matched in the breadth-first search in each respective sequence.

38. The method of claim 37, wherein a series of local mass-based alignments are made until the entire *de novo* sequence has been accounted for by the sequence in the sequence database in the mass-based alignments.

39. The method of claim 38, wherein a maximum number of consecutive modification sites are performed.

40. The method of claim 39, wherein the maximum number of consecutive modification sites is 1-or 2 local mass-based alignments in length.

41. The method of claim 39, wherein the modification information about modifications is cataloged in a modification catalog.

42. The method of claim 41, wherein the modification information includes at least one of, molecular mass of the modification, a specific molecules where the modification occurs, a frequency of

occurrence of the modification at those molecules, wherein the frequency of occurrence is the estimated frequency in nature or a frequency as a sample preparation artifact, a mass object for the modification, which represents the additional mass of the modification to the *de novo* sequence at those molecules, and the name of the modification, and a modification score for the modification.

43. The method of claim 42, wherein a modification is selected from, an *in vivo* or *in vitro* protein, a peptide modification, and an amino acid substitution.

44. The method of claim 43, further comprising: ranking the modifications, wherein the ranking is based on their frequency of occurrence.

45. The method of claim 44, further comprising: identifying a most probable modification in the modification site from the modification catalog by matching elements to elements in modifications in the modification catalog that are selected from at least one of, the mass difference, the molecules in the sequence database in the modification site, and the ranking of the modification in the modifications catalog.

46. The method of claim 45, wherein the mass difference is the difference between the sum of the molecular masses of the consecutive mass objects in the sequence in the sequence database and the sum of the molecular masses of the consecutive mass objects in a *de novo* sequence in a local mass-based alignment.

47. The method of claim 45, wherein the mass object of an identified modification is inserted into the in the mass-based alignment, which creates a mass match between the *de novo* sequence and the sequence in the sequence database.

48. The method of claim 38, further comprising: computing a match score of the mass-based alignment, the match score being a measure of how well the sequence in the sequence database matches the *de novo* sequence.

49. The method of claim 48, wherein a match score is generated from the linear combination of local alignment scores from the series of local mass-based alignments.

50. The method of claim 49, wherein each of a series of consecutive local mass-based alignments receives a score and is classified.

51. The method of claim 50, wherein each local alignment score is generated using a substitution matrix, depending on whether the local alignment is a mass match, a modification site, or a gap match.

52. The method of claim 51, wherein the substitution matrix contains substitution matrix score of least one molecule.

53. The method of claim 52, wherein the substitution matrix identity score is a substitution matrix score between a molecule and itself.

54. The method of claim 53, wherein the substitution matrix substitution score is a substitution matrix score between a molecule and a different molecule.

55. The method of claim 54, wherein the substitution matrix score is the log of the odds score of an identity of a molecule or a substitution between two molecules.

56. The method of claim 52, wherein the local alignment score for a mass match is the average value of the substitution matrix identity scores for all of the molecules in the sequence in the sequence database matched in the local alignment.

57. The method of claim 56, wherein if at least one of the molecules has been modified by a modification, the substitution matrix score for each modified molecule is the modification score for that modification.

58. The method of claim 52, wherein if the local mass-based alignment is a match between only one mass object from the sequence in the sequence database, and only one mass object from the *de novo* sequence, and that those mass objects represent single molecules, then the local alignment score for a substitution is the substitution matrix substitution score between the molecule in the sequence in the sequence database and the molecule in the *de novo* sequence.

59. The method of claim 52, wherein the local alignment score for a substitution is the number of molecules in the substitution in the sequence in the sequence database multiplied by the average value of the substitution matrix substitution scores.

60. The method of claim 52, wherein the local alignment score for a gap match is the number of molecules in the gap match in the *de novo* sequence multiplied by the average value of the substitution matrix substitution scores.

61. The method of claim 48, wherein if the termini of the *de novo* sequence are expected to be specific molecules, the match score is increased if the termini of the mass-based alignment match the expected specific molecules.

62. The method of claim 48, wherein if the termini of the *de novo* sequence are expected to be specific molecules, the match score is decreased if the termini of the mass-based alignment do not match the expected specific molecules, or if expected specific molecules are present inside the mass-based alignment.

63. The method of claim 1, further comprising utilizing an approach that interprets matches between sequences in the sequence database and *de novo* sequences, which have been scored by a match score, as an identified macromolecule list and assigns a macromolecule score to each sequence in the identified macromolecule list.

64. The method of claim 63, wherein the match score is a measure of how well the sequence in the sequence database matches the *de novo* sequence.

65. The method of claim 64, wherein *de novo* sequences that match at least one sequence in the sequence database are classified as either discriminating *de novo* sequences or non-discriminating *de novo* sequences, the *de novo* sequences are inserted into a *de novo* sequence list, and the *de novo* sequences in the *de novo* sequence list are ranked by their delta scores.

66. The method of claim 65, wherein the delta score is computed for the *de novo* sequence as the difference between the match scores of the first and second matches to sequences in the sequence database for that *de novo* sequence. If that *de novo* sequence only matches one sequence in the sequence database, the delta score is the match score for that match.

67. The method of claim 66, wherein discriminating *de novo* sequences have a delta score greater than or equal to the delta score threshold and non-discriminating *de novo* sequences have a delta score less than the delta score threshold.

68. The method of claim 67, wherein the delta score threshold for the *de novo* sequence is between 0% and 25% of the match score of the highest scoring match between a sequence in the sequence database and that *de novo* sequence.

69. The method of claim 67, All matches between a sequence in the sequence database and a *de novo* sequence with match scores less than the match score of the highest scoring match between a sequence in the sequence database and that *de novo* sequence minus the delta score threshold are discarded.

70. The method of claim 60, wherein the sequence in the sequence database, which matches best to the discriminating *de novo* sequence in the *de novo* sequence list with the greatest delta score, is added to the identified macromolecule list. This *de novo* sequence is then moved from the *de novo* sequence list to that sequence.

71. The method of claim 70, wherein all non-discriminating *de novo* sequences in the *de novo* sequence list that match to that sequence in the identified macromolecule list are moved from the *de novo* sequence list to that sequence.

72. The method of claim 71, wherein the process of 1 is repeated until all discriminating *de novo* sequences in the *de novo* sequence list are removed from the *de novo* sequence list.

73. The method of claim 72, wherein all sequences in the sequence database that match to non-discriminating *de novo* sequences still in the *de novo* sequence list are added to the identified macromolecule list, and the non-discriminating *de novo* sequences still in the *de novo* sequence list are moved to those sequences.

74. The method of claim 73, wherein a macromolecule score is generated for every sequence in the identified macromolecule list.

75. The method of claim 74, wherein the macromolecule score is a linear combination of the *de novo* macromolecule scores of the *de novo* sequences that have been assigned to that sequence.

76. The method of claim 64, a new sequence database is generated containing only the sequences in the sequence database that are listed in the identified macromolecule list.

77. The method of claim 76, wherein *de novo* sequences that do not match any sequence in the original sequence database are re-analyzed by calculating a mass-based alignment between each *de novo* sequence in question and sequences in the new sequence database, as described in claim 1 in a way that the search space explored by the mass-based alignment algorithm is increased.

78. The method of claim 77, further comprising: decreasing the specified length of tags.

79. The method of claim 77, further comprising: increasing the search depth.

80. The method of claim 77, further comprising: increasing the maximum number of consecutive substitutions.

81. The method of claim 64, wherein *de novo* sequences that do not match any sequence in the original sequence database are re-analyzed by calculating a mass-based alignment between each *de novo* sequence in question and sequences in a different sequence database, as described in claim 1.

82. A method for identifying sequences of molecules and sequence modifications from mass spectrometry data comprising:

a. producing at least one *de novo* sequence from mass spectrometry data of sequences of molecules,

b. calculating at least one mass-based alignment between each *de novo* sequence and sequences in a sequence database, wherein the molecular masses of molecules in the *de novo* sequence are compared to molecular masses of molecules in each sequence in the sequence database,

c. interpreting mass differences of modification sites between the sequence in the sequence database and the *de novo* sequence that have been identified by the mass-based alignment as modifications identified in a modification catalog, and

d. calculating at least one match score for the mass-based alignment that provides an indication of matching between the sequence in the sequence database and the *de novo* sequence.

83. The method of claim 82, further comprising: identifying sequences in the sequence database from mass-based alignments in response to the match scores.

84. The method of claim 83, further comprising: grouping identifications of sequences in the sequence database from at least one *de novo* sequence into an identified macromolecule list that agrees with the de novo sequencing results.

85. A computer readable medium having stored thereon instructions which, when executed by a processor, cause the processor to perform:

a. executing a first application that produces at least one *de novo* sequence from mass spectrometry data of sequences of molecules,

b. executing a second application that calculates at least one mass-based alignment between each *de novo* sequence and sequences in a sequence database, wherein the molecular masses of molecules in the *de novo* sequence are compared to molecular masses of molecules in each sequence in the sequence database,

c. executes a third program that interprets mass differences of modification sites between the sequence in the sequence database and the *de novo* sequence that have been identified by the mass-based alignment as modifications identified in a modification catalog, and

d. executes a fourth program that calculates at least one match score for the mass-based alignment that provides an indication of matching between the sequence in the sequence database and the *de novo* sequence.

86. The computer readable medium of claim 85, wherein the processor further executes a fifth program that identifies sequences in the sequence database from mass-based alignments in response to the match scores.

87. The computer readable medium of claim 86, wherein the processor further executes a sixth program that groups identifications of sequences in the sequence database from at least one *de novo* sequence into an identified macromolecule list that agrees with the de novo sequencing results.

88. A computer based system that implements identification sequences of molecules and sequence modifications from mass spectrometry data, comprising at least a first processor that executes one or more programs that:

a. produces at least one *de novo* sequence from mass spectrometry data of sequences of molecules,

b. executing a second application that calculates at least one mass-based alignment between each *de novo* sequence and sequences in a sequence database, wherein the molecular masses of molecules in the *de novo* sequence are compared to molecular masses of molecules in each sequence in the sequence database, and

c. executes a third program that interprets mass differences of modification sites between the sequence in the sequence database and the *de novo* sequence that have been identified by the mass-based alignment as modifications identified in a modification catalog, and

d. executes a fourth program that calculates at least one match score for the mass-based alignment that provides an indication of matching between the sequence in the sequence database and the *de novo* sequence.

89. The computer based system of claim 88, wherein at least a first processor executes one or more programs that identifies sequences in the sequence database from mass-based alignments in response to the match scores.

90. The computer based system of claim 89, wherein at least a first processor executes one or more programs that groups identifications of sequences in the sequence database from at least one *de novo* sequence into an identified macromolecule list that agrees with the de novo sequencing results.